Validation of Empirically Derived Rating Scales for a Story Retelling Speaking Test

doi: 10.1080/15434303.2013.824973

Language Assessment Quarterly, 10:1-21, 2013
DOI: 10.1080/15434303.2013.824973

Validation of Empirically Derived Rating Scales for a Story Retelling Speaking Test

Akiyo Hirai
University of Tsukuba, Tsukuba, Japan

Rie Koizumi
Juntendo University, Chiba, Japan

In recognition of the rating scale as a crucial tool of performance assessment, this study aims to establish a rating scale suitable for a Story Retelling Speaking Test (SRST), which is a semi-direct test of speaking ability in English as a foreign language (EFL) for classroom use. To identify an appropriate scale, three rating scales, all of which have been designed to have diagnostic functions, were developed for the SRST and compared in terms of their reliability, validity, and practicality. The three scales were: (a) an empirically derived, binary-choice, boundary-definition (called EBB1) scale, which has four criteria (Communicative Efficiency, Content, Grammar & Vocabulary, and Pronunciation); (b) an EBB2 scale that was modified from the EBB1 scale and has three criteria (Communicative Efficiency, Grammar & Vocabulary, and Pronunciation); and (c) a multiple-trait (MT) scale that was modified from the EBB2 but has a conventional analytic scale format. The results of the comparison revealed that the EBB2 was the most reliable and valid measure for assessing speech performance in the context of story retelling. However, the MT was shown to be the most practical, while the EBB2 permits more careful scoring, which suggests the influence of the rating scale format on test qualities.

INTRODUCTION

There is a growing awareness of teachers' responsibility to assess their students' learning and also of the impact that assessment has on learning (e.g., Hill & McNamara, 2012). Thus, this study focuses on the development of a rating scale for classroom assessment. Among a variety of factors affecting the assessment of speaking performance, such as raters, rating scales, interlocutors, elicitation tasks, and test-taker

proficiency (Fulcher, 2003; Luoma, 2004), rating scales have been especially scrutinized because they "provide an operational definition of a linguistic construct" (Fulcher, 2003, p. 89) and should properly reflect a construct, or what we intend to assess (McNamara, 1996). In this regard, developing valid and reliable rating scales is of great importance in successfully assessing speaking performance.

In addition, one of the greatest challenges in performance assessment is practicality. Rating procedures often take a large amount of time by requiring teachers to listen to student performances individually. Moreover, the use of commercially available speaking tests imposes a financial burden on the students. For that reason, classroom teachers are reluctant to use such tests to assess classes of about 40 students (e.g., Honda, 2007). In this regard, time- and cost-effectiveness are particularly important for tools used in practical classroom assessment.

The speaking test for which the scale is being created is the Story Retelling Speaking Test (SRST), a user-friendly, semi-direct speaking test that uses an integrated reading-to-retell task developed for classroom use by the authors (see the "Procedure of the SRST" section and Appendix A; Hirai & Koizumi, 2009). On the basis of the results of the questionnaire used in the study, the SRST was found to be suitable for classroom use and to give positive washback to EFL students with low to intermediate proficiency levels. This positive washback can be attributed to three factors. First, the test elicits extended production of story retelling and opinion statement in English, which Japanese students are known to be poor at producing (National Institute for Educational Policy Research of Japan, 2007). Second, even low-proficiency students are encouraged to produce the target language using phrases first presented in a text (i.e., story) and to express their opinions about the story. Third, conducting a reading-to-retell activity such as the SRST is encouraged by the Ministry of Education, Culture, Sports, Science and Technology (n.d.), which specifies in the Course of Study for Upper Secondary School that "instruction on speaking and writing should be conducted more effectively through integration with listening and reading activities" (p. 2). Moreover, we can expect that if performance in the reading-to-retell task is evaluated, students may approach it more seriously, thereby working to improve their speaking skills.

From a teacher's perspective, the preparation and administration of the test are easy because all the teacher has to do is to prepare a short text for the students to retell, and give the test to all the students at one time during a class. In addition, providing feedback to students is relatively easy because the original text can serve as a model answer.

Rating Scales for Classroom Use

In constructing a rating scale for classroom use, we have kept at least two points in mind. One is that the rating scale should be specific to the task (i.e., the SRST); the other is that it should provide some form of diagnostic information for students. Concerning the first point, Turner and Upshur (1996) pointed out the following problem: Descriptors of a scale created based on an impression or general theory of the development of language abilities did not conform to the teaching objectives in their study, and only a portion of the theory-based scale was applicable to the classroom situation. Thus, more empirically based rating scales that are derived from samples of test performances are encouraged (Knoch, 2007, 2009). As for the second point, in order to give diagnostic feedback to the students and enable them to recognize which areas they should improve, a rating scale must have multiple assessment criteria.

The following two types of empirically based rating scales may meet these criteria. One is an empirically derived, binary-choice, boundary-definition (EBB) scale (e.g., Upshur & Turner, 1995); the other is a multiple-trait (MT) scale (e.g., Hamp-Lyons, 1991).

The EBB rating scale measures several levels of a single trait or multiple traits; thus, it can provide either a single score or multiple scores for each speech sample. As shown in Appendix B, the EBB scale is unique in that it is composed of a hierarchical set of articulated binary questions or descriptions. The scale divides adjacent score levels, and its descriptors are defined on the basis of a selection of actual performances of a task. Thus, raters observe participants' performances and make a series of yes-or-no choices to arrive at a final score.

Turner and Upshur (1996) reported the results of EBB scales created and implemented by a team of teachers as follows. First, no raters reported difficulty in scoring speech samples using the scales. Rating reliability by the Facets analysis was satisfactory. Second, the raters found that the scale reflected the aim of the task more specifically than theory-based rating scales did. Third, EBB scores were distributed over all six scoring levels, which implies that the scale discriminated learners' speaking abilities well. Fourth, the scoring of speaking ability was time-efficient. Raters were able to rate 15 to 20 speech samples per hour.

These results suggest that the EBB scale is a practical and reliable method. However, Fulcher (2003) pointed out that, since the EBB scale is developed specifically in reference to a certain task, it is difficult to use it to generalize a test result across different tasks (see also Bachman & Savignon, 1986). At the same time, he considered this scale to be preferable, writing that, "despite this problem [i.e., the lack of

generalizability], the explicitness of the design methodology for EBBs is impressive, and their usefulness in pedagogic settings is attractive" (p. 107). In this regard, an EBB scale can be tailored to the assessment objectives of a particular classroom and can measure the intended constructs precisely.

Another type of empirical rating scale is the MT scale, which focuses on multiple dimensions of a performance (see an example in Appendix D). Depending on which features of the speech sample the scores represent, MT scales can be either task-specific or generalized across a range of task types, and performance is evaluated in terms of several traits (i.e., criteria), each with several levels. Given that the criteria in MT rubrics focus on the specific features of performance in a given task or tasks, they can reflect intended constructs more sensitively; thus, they provide more specific diagnostic information (Hamp-Lyons, 1991).

On the surface, MT scales look like analytic scales. However, TM scales focus on specific features of given tasks, while analytic scales usually evaluate the more generic dimensions of language production. Thus, as well as EBB scales, MT scales can obtain construct and content validity of criterion-referenced assessments, and benefit students by providing specific feedback about the students' strengths and weaknesses in the task with MT rubrics (Center for Advanced Research on Language Acquisition, 2009). MT scales also share disadvantages with EBB scales in that their task-specific rubrics have limited use for other tasks unless they are modified appropriately.

A Preliminary Study of the EBB Scale

Considering the advantages of EBB scales, as described above, in a preliminary study, we (Hirai & Koizumi, 2008) attempted to develop an EBB scale with multiple criteria for scoring the SRST, referring to the Course of Study (Ministry of Education, Culture, Sports, Science and Technology, n.d.), Upshur and Turner's (1995) notions of communicative effectiveness and grammatical accuracy, and the Foreign Service Institute scale's concepts of fluency, grammar, vocabulary, and accent (Hughes, 2003). To cover the intended linguistic aspects of speaking ability as mentioned in the Course of Study's instructional goals, the EBB scale we created (hereafter called EBB1; see Appendix B) consisted of four trait criteria with five hierarchical levels each: Communicative Efficiency, Content, Grammar & Vocabulary, and Pronunciation. Communicative Efficiency measures fluency, while Content measures coherency, elaboration, and sufficiency of both the retold story and student opinion. Grammar & Vocabulary focuses on grammatical accuracy and vocabulary use. Finally, the

Pronunciation criterion includes not only the accuracy of pronunciation and stress, but also accuracy of prosody such as intonation and rhythm.

The construction of score descriptors for these criteria was carried out using Turner and Upshur (1996) as a guide. We first classified students according to their performance levels, examined prominent features that separated the levels by inspecting both speech and transcribed performances by students, and independently created the descriptor of each criterion. Next, through discussion, we identified five hierarchical levels for distinct separation of the performances. Thus, with a descriptor at each distinct separation point, the final scale was shaped like a family tree (see Appendix B).

The EBB1 was then used to score the students' performances in the SRST. The results revealed that, with the exception of the Grammar & Vocabulary criterion, the three criteria worked well. Moreover, analysis using the generalizability theory showed that only two stories (i.e., two tasks) were sufficient to achieve a reliability (phi coefficient) of .70 or above, which was determined to be the acceptable level for a relatively low-stakes speaking test (e.g., Lado, 1964; Nunnally, 1978; Santos, 1999). However, the Grammar & Vocabulary criterion required three stories to achieve such a level of reliability. Thus, the descriptors of the Grammar & Vocabulary criterion needed some modification.

In regard to the validity of the EBB1, the correlation between Communicative Efficiency and Content criteria was high at over .80, suggesting that the two shared a large proportion of variances. Close inspection of the EBB descriptors revealed two points. First, when students spoke fluently for two minutes, thereby increasing their production, they could cover wider aspects of content. Second, because the stories were easy and the topics were familiar, hardly any of the students had trouble comprehending them (Koizumi & Hirai, 2010). This suggested that judgment about "content errors," a component of the Content criterion, was not required. For these reasons, the two criteria were considered to assess similar constructs and, therefore, they could be combined into a single criterion.

Purpose of the Present Study

Since the EBB1 in the preliminary study was found to be problematic, the present study aims to modify it as appropriate and examine whether the modified EBB (called EBB2) scale can resolve the problems of the EBB1 (see Appendix C). This study also examines the MT scale type in comparison to the EBB scale type because, as seen in the literature review, the MT scale is another promising empirically constructed scale and its format is much more commonly used. Thus, we created an MT scale based on the

EBB2. The MT scale uses basically the same rating descriptors as the EBB2, but it looks like a traditional analytic scale (see Appendix D). Therefore, as illustrated in Figure 1, the present study is an extension of the preliminary study and compares the EBB1, EBB2, and MT in order to determine which is best suited to the SRST.

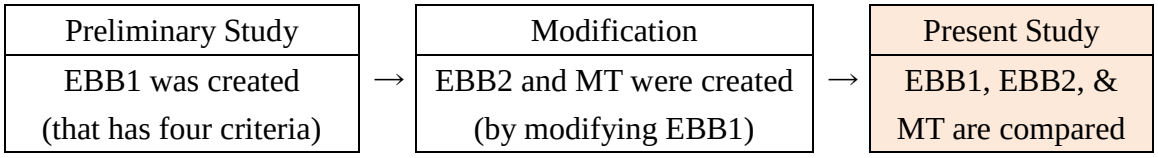


Figure 1. Connection between preliminary and present studies.

If practical and useful rating scales are available for classroom speaking tasks, teachers might be encouraged to make more consistent use of speaking tests in the classroom. Further, there has been little research so far on speaking tests and rating scales that are suitable for classroom use or how different types of rubrics affect rater performance from the standpoints of validity, reliability, and practicality (e.g., Barkaoui, 2007, 2010; Knoch, 2007, 2009). Thus, the present study fills the gap in these areas for an enhancement of classroom assessment of EFL speaking ability in Japan.

In investigating the quality of rating scales, a variety of measures have been used (e.g., Knoch, 2007, 2009; Lumley & Brown, 2005; McNamara, 1996). Among them, Knoch (2007, 2009) recommended comparing similar scales using a many-faceted Rasch analysis. She investigated whether an empirically constructed scale functions better than a traditional-type scale in writing assessment, and provided evidence of different outcomes for the two scales. Since Knoch's (2007, 2009) procedures appear to be thorough and the purpose of her study parallels that of our own, we will refer to her analytical framework in the present research. Thus, this study poses the following four research questions (RQs).

RQ1. How many stories are needed to achieve sufficient reliability when the EBB1, EBB2, and MT rating scales are used for scoring?

RQ2. To what extent do the individual trait criteria of the EBB1, EBB2, and MT scales differ in terms of (a) the discrimination of the rating scale, (b) rater separation, (c) rater reliability, (d) variability in the ratings, and (e) rating scale properties?

RQ3. To what extent are the individual trait criteria of the EBB1, EBB2, and MT scales related to each other?

RQ4. What are the raters' perceptions of the practicality of the EBB1, EBB2, and MT scales?

METHOD

Participants

Speech performances were obtained from 48 Japanese university students. Among them, 44 were undergraduate students majoring in English or Engineering. Their English proficiency levels were regarded as beginner to intermediate. A preliminary study had used productive vocabulary tests to measure the proficiency of students who were taking the same levels of classes as the participants in the main study (Hirai & Koizumi, 2009). The remaining four participants were graduate students majoring in English; they were generally more proficient than the former group.

SRST Procedure

All students in the study took the SRST, which is a tape-mediated test. First, they silently read a story for two minutes, and then answered three comprehension questions orally (see Appendix A). Next, they retold as much of the story as possible in two minutes, looking only at four keywords to help them recall it. They were also told to include their opinions about the story content. Although the students were not instructed about what type of opinions they should include, most of them made comments on the main characters' behaviors, on what they would do if they were in the main characters' shoes, and on lessons they had learned from the story (Koizumi & Hirai, 2010).

This procedure was repeated for each of the four stories, and the order of administering stories was varied for all the students to counterbalance the order effect. The retelling performances were recorded for scoring with the EBB1 in the preliminary study and the EBB2 and MT in the present study.

The four stories used in the test were two 100-word and two 150-word texts adopted from past Eiken Grade 3 and 4 tests (Society for Testing English Proficiency, 2012), which are the most widely used standardized tests in Japan. The difficulty level of the text was nearly the same for the four stories (e.g., Flesch-Kincaid Grade Levels: 4.1, 4.5, 4.6, and 5.3, respectively). Stories of different lengths and the same difficulty were found not to substantially affect students' story retelling, and this suggests that the SRST does not significantly assess short-term memory capacity (Koizumi & Hirai, 2010).

Rating Scales

EBB1. As mentioned previously, the EBB1 was created in our previous study and contained the following criteria: (a) Communicative Efficiency, (b) Content, (c) Grammar & Vocabulary, and (d) Pronunciation. Each criterion has five hierarchical levels (Appendix B).

EBB2. On the basis of the results of the preliminary study, we made two major modifications to the EBB1 to create the EBB2 (see Appendix C).

The first revision was made to the EBB1 Grammar & Vocabulary criterion. In the preliminary study, it was found that this criterion was unstable when compared with the other three criteria. One cause of this instability was that participants who spoke less tended to make fewer errors, and consequently attained a high score. Thus, in the modified version, we gave the lowest rating when the utterances contained only a few sentences by adding the words “or with few sentences" in the third descriptor (see Appendices B and C), assuming that the grammatical knowledge of students who spoke little was so lacking that they were unable to produce a series of sentences.

Next, we combined the Communicative Efficiency and Content criteria in the EBB1 into one “Communicative Efficiency” criterion, for the reasons mentioned in the section titled “A Preliminary Study of the EBB Scale.” Thus, the number of criteria was reduced from four to three, which may have made scoring easier for the raters, given that in the preliminary study some raters mentioned that it was difficult to assign scores according to all four criteria by listening once to a performance that lasted only two minutes.

With these two major modifications, and other small changes in the wording of the descriptors, we created the EBB2, which consists of three criteria with five hierarchical levels each.

MT scale. We changed the EBB2 into an analytic format and called it an MT, as it is empirically based and has multiple criteria. Therefore, apart from its format, the MT is identical to the EBB2, consisting of three criteria with five hierarchical levels each (see Appendix D).

Questionnaire

A questionnaire, which was the same one used in the preliminary study, was given to the raters who had scored the student performances on both the EBB2 and the MT. It consisted of five questions about the practicality of the two scales, specifically, the ease

and efficiency of their use. The results of the questionnaire were then compared with the results obtained for the EBB1 in the previous study.

Rating

EBB1 rating. In 2008, six raters (two university teachers and four graduate students majoring in English language education) attended an approximately two- to three-hour training session. First, the raters were given an explanation of certain words in the descriptors that were difficult to interpret, such as prosody, stress, and accent. Then, they listened to several benchmark performances and scored them until they reached a consensus. After the training session, each rater scored 10 or 11 participant performances. Thus, in total, 272 rated performances (i.e., 6 raters x [10 or 11 examinees] x 4 stories) were analyzed. Of these, approximately 80 performances were scored by two raters to measure inter-rater reliability.

EBB2 and MT ratings. In 2009, nine raters participated in the main study: the same two teachers and three graduate students who had participated in the preliminary study, along with four newly recruited graduate students majoring in English language education. The approach to rater training was the same as before. They practiced on the same samples, using the EBB2 and MT, one at a time. After training, all raters used both the EBB2 and MT; half used the EBB2 first, and the other half used the MT first in order to avoid the order effect. In all, the EBB2 was used to rate 272 performances (i.e., 9 raters x [7 or 8 examinees] x 4 stories), and the MT was used to rate 248 performances (i.e., 9 raters x [6 or 7 examinees] x 4 stories). Among them, 80 performances were rated by two raters using the EBB2, whereas 56 performances were rated by two raters using the MT to measure inter-rater reliability.

The raters belonged to the same population for which the scale is intended to be used (i.e., current and future English teachers). In addition, although the EBB1 rating was made for a different period, it is comparable to the EBB2 and MT ratings because the rating samples and procedures were the same. Additionally, the long period between the two rating sessions was presumed to reduce the effect of familiarity on rating performance and questionnaire responses, although some of the raters participated in both studies.

Data Analysis

To examine RQ1 regarding how many stories are needed to achieve sufficient reliability by using the EBB1, EBB2, and MT rating scales, we adopted the multivariate

generalizability theory (Brennan, 2001), utilizing the mGENOVA program (Center for Advanced Studies in Measurement and Assessment, 2009). All the criteria of a rating scale were analyzed with a crossed one-facet (i.e., story) design to obtain reliability for individual criteria. We did not include the rater facet in the design, even though this may limit the generalizability of the results. There were two reasons for this. Firstly, a many-faceted Rasch analysis can investigate the inter-rater reliability and relevant aspects. Secondly, we intended to examine RQ1 under the same conditions in which the SRST is assessed by a single teacher, given that this is how classroom assessment is usually conducted.

To address RQ2, whether the EBB1, EBB2, and MT function properly with regard to reliability and validity, a many-faceted Rasch measurement program, Facets (Linacre, 2009a), was used. It was selected for its ability to analyze the four aspects (i.e., facets) of participants, raters, stories, and rating criteria with the same measurement scale. We applied the partial credit model with the Facets program to analyze individual rating criteria, and the rating scale model to analyze the overall scale.

Regarding RQ3, a multitrait-multimethod (MTMM) approach and principal component analysis (PCA) were employed to examine whether each rating scale could measure our intended construct and how the scales related to each other. The mean ratings were used when two raters rated the same participants for RQ1 analysis, while RQ3 analysis used the mean ratings of four stories and two raters, when available.

Finally, in order to answer RQ4, we compared the practicality of the three scales by administering the same questionnaire that had been used in the preliminary study for the EBB1.

This study employed multiple methods to examine scale quality comprehensively, by focusing on the following in RQ1 to RQ4, respectively: task reliability, rater reliability, validity, and practicality.

RESULTS

Reliability of the Scales

The reliability coefficients of the scores rated by each scale were first compared using a generalizability (G) study. We estimated the following three sources of variability: (a) differences among the objects of measurement (i.e., the persons in the study), (b) differences in item difficulty (i.e., stories), and (c) the residual, including the person-by-story interaction and random errors (i.e., *ps, e*).

Table 1 presents the variance components for persons (*p*) in the second column, variance components for stories (*s*) in the third column, and variance components for

the residual (*ps, e*) in the fourth column. It suggests that the large variability of the scores as measured by the EBB1, EBB2, and MT was explained by the variance component for persons. For example, the variance component for persons in the Communicative Efficiency criterion was 74.09%, 62.07%, and 60.28% in the EBB1, EBB2, and MT, respectively. This is a favorable result because variability in the measurement substantially reflects differences in the participants' intended abilities. On the other hand, the variance component for stories in all criteria for the three scales was marginal, ranging from 0.00% to 6.48%, which suggests that the stories had similar levels of difficulty.

Insert Table 1 about here.

Regarding the variance component for the residual, that of the EBB2 Grammar & Vocabulary was much smaller (27.26%) than those of the EBB1 and MT (54.05% and 48.71%). The smaller variance revealed in the EBB2 criterion was found to be superior because the relative ordering of participants differed little across stories. This finding implies that the problems with the EBB1 Grammar & Vocabulary criterion were resolved in the EBB2.

The decision (D) studies allow us to examine how the reliability coefficient of each criterion changes depending on the number of stories. Because the rating scales are criterion-referenced, phi coefficients (Φ), measurements for an absolute decision were interpreted, instead of generalizability (G) coefficients that are for a relative decision.

By observing the phi coefficients over .70 (shown in boldface in Table 2), which was judged to be a sufficient level of reliability for a classroom speaking test, we can see that the three criteria of the EBB2 would necessitate only one or two stories in the SRST. In contrast, the Grammar & Vocabulary criteria in both the EBB1 and MT would require three stories in order to achieve a reliable rating. These results suggest that EBB2 scores tend to be slightly more generalizable than EBB1 scores, and that two stories are sufficient to produce reliable scores with the EBB2.

Insert Table 2 about here.

Functions of the Scales

Next, we investigated whether the EBB1, EBB2, and MT function properly for individual rating criteria and for each scale as a whole, in terms of the following points:

(a) discrimination of the rating scale, (b) rater separation ratio, (c) rater reliability, (d) variability in ratings, and (e) rating scale properties.

The discrimination of the rating scale (a) can be measured by the person separation ratio. A larger separation ratio is positive because it indicates that the rating scale can discriminate between good and poor performances to a greater degree. On the other hand, a smaller rater separation ratio (b) is a positive result because it indicates that the rater differences are smaller in terms of severity and leniency. For the third indicator of rater reliability (c), the Facets program provides two measures. One is the rater point biserial correlation, or a measure of how similarly the raters rank the candidates. The other is the percentage of exact rater agreement, which indicates the number of times a rater gave exactly the same score as another rater in the sample. Higher values for rater point biserial correlation and exact agreement ratio indicate higher inter-rater reliability.

In terms of variability in the ratings (d), if raters score performances either inconsistently or overly consistently (i.e., they tend to overuse middle band levels), they can be judged as misfit or overfit in rater infit mean square values. Facets manual (Linacre, 2009b) explains that infit mean square values from 0.50 to 1.50 are productive for measurement. Values higher than 1.5 indicate significant misfit and might be a threat to validity, while values lower than 0.5 indicate significant overfit but do not degrade the measurement system. According to Bond and Fox (2007), a reasonable value range for raters or judges where agreement is encouraged is suggested to be between 0.4 and 1.2. Considering that our scale is intended for classroom assessment and overfit raters do not have a serious impact when rating a relatively homogenous group of students, we set moderate cut-off values: values higher than 1.4 as significant misfit and values lower than 0.4 as significant overfit.

The last indicator is rating scale properties (e). Ideal conditions are for the level of a rating criterion to increase steadily in difficulty from Level 1 to Level 5, with performances normally distributed over the five levels. According to Linacre (2009b), for a reliable estimation, each band level needs to include at least 10 observations (i.e., ratings). We report the average difficulty in each band level in logit value, followed by the number (Count) and percentage (%) of ratings.

Communicative Efficiency. The EBB1 Content criterion is shown in Table 3 because it was included in the Communicative Efficiency criterion of the other two scales. The results of the rater separation ratio and the exact agreement ratio improved from the EBB1 to the EBB2, and there were no misfit raters for the EBB2. On the other hand, the MT did not show much improvement from the EBB1 because the point

biserial correction for the rater and exact agreement ratio were the lowest among the scales. In addition, there were two overfit raters for the MT, which means they tended to score only in the same middle band levels. In terms of the scale properties, participants were more normally distributed in the EBB2 and MT than in the EBB1. The EBB1 was negatively skewed, primarily because Level 4 was most selected (42%). At the first descriptor, "Coherent story retold with no long pauses," raters generally chose "Yes" and selected Level 4 or 5; at the next descriptor separating Levels 4 and 5, "with little hesitation and with few self-corrections," they tended to choose Level 4. In other words, once the first decision was made, they often selected Level 4. In fact, as Rater 5 in the questionnaire (see Appendix E) pointed out, if a rater makes a wrong decision at the early branching nodes (i.e., decision points) of an EBB scale when choosing either "Yes" or "No," it can make a significant difference in the final score. This problem was successfully solved by providing more descriptors (i.e., decision points) before reaching Level 4 in the EBB2.

The EBB2 scale would be better for another reason. Since "fluency" seems to be an overarching feature of "pausing," "hesitation," and "self-correction" in the Communicative Efficiency criterion, it would make more sense to judge test-takers' "fluency" feature first as indicated in the first node of the EBB2 rather than their "pause" feature in the EBB1. For all these reasons, the EBB2 functioned better and was the most reliable of the three scales.

Insert Table 3 about here.

Grammar & Vocabulary. Since modifications were made to the last two decision points (i.e., nodes) of EBB2, we explained this in the rater training session in order to get raters to notice these points when scoring. As a result of the revisions and the rater training, the problem with the Grammar & Vocabulary section of the EBB1 seemed to be solved in the EBB2, in which the rating distribution was normally distributed in the scale properties with a better person separation ratio (2.71) and higher rater point biserial correlation (.47; see Table 4). Fewer than 10 performances rated as Level 2 on the EBB1 increased to 72 performances (26%) with the EBB2, owing to the addition of the words "with some prominent grammatical and lexical errors" at the fourth descriptor. Further, performances that contain fewer errors due to participants' limited production were rated as Level 1 by adding the words "or with few sentences" in the EBB2 descriptor. This is evidenced by more ratings (15%) being assigned to Level 1 on the EBB2 than on the EBB1 (11%).

Of the three scales, the most problematic is the MT, because it has the lowest exact agreement ratio (41.1%) and there were two raters who displayed misfit values. These results indicate that raters tended to have difficulty in rating consistently when using the MT. This difficulty may be attributed to the format of the scale. As implied by Rater 7 (see Appendix E), exposure to all the score descriptors at once in the MT might have created too much of a cognitive demand on the raters, which may have led to fluctuating ratings across the five levels. Based on these results, the EBB2 appears to be the best choice among the three scales.

Insert Table 4 about here.

Pronunciation. As shown in Table 5, a noticeable improvement from the EBB1 to the EBB2 was also observed in that the EBB2 has the smallest rater separation ratio (1.55) and the highest exact agreement ratio (66.2%) of the three scales. In addition, nearly half (43%) of the ratings categorized as Level 4 on the EBB1 were reasonably reduced to 17% on the EBB2, in which the meaning of "prominent prosodic errors" was clarified with the addition of the words "such as word level stress" (see Appendix C).

Furthermore, as indicated with logit values, the difficulties of Levels 1 and 2 (-2.39 and -1.05) of the EBB2 were separated more clearly than those (-0.74 and -0.72) of the EBB1. This may have been achieved by different weightings in "strong accent" and "frequent prosodic errors" in the EBB2 and EBB1 and subsequent rater training sessions. In the EBB2, test-takers obtain the lowest score of Level l if their performances contain "frequent prosodic errors" regardless of their accent as shown in the fourth node. In the EBB1, test takers gain Level 1 if their performances contain "frequent prosodic errors" and "strong accent." This difference indicates that the EBB2 attaches a lighter weight to "accent" and a heavier weight to "frequent prosodic errors" than the EBB1 does. This modification reflects our belief that we should first eliminate noticeable pronunciation errors.

In contrast, using the MT, two raters had overly inconsistent results, as indicated by two misfit raters. In addition, only three participant performances (i.e., 1%) were scored at Level 5. In regard to the one overfit rater in each scale, it might have been difficult for the raters to separate a relatively homogeneous group of test-takers' pronunciation ability into five band levels no matter how the rating descriptors were modified or how clearly key terms such as "frequent prosodic errors" and "strong accent" were elaborated in the rater training sessions.

In light of these results, it can be said that the Pronunciation criterion of the EBB2 functioned the most successfully.

Insert Table 5 about here.

Overall ratings. The criteria of all the scales analyzed together are shown in Table 6. The person separation ratio of the EBB2 was the highest (3.94), suggesting that the EBB2 has the greatest discriminatory power among the three scales. In addition, the difficulty values of the five levels were spread more widely and equally (from logit values of -2.23 to 3.24) than those in the other two scales. In terms of rater separation and reliability in ratings, the EBB2 was again found to be the most reliable because its rater separation was the smallest, and both the rater point biserial correlation and the exact agreement ratio were the highest. There was, however, one rater who displayed a misfit value, which may indicate a need for further rater training. On the other hand, the MT was found to be most problematic in that ratings of three raters were identified as misfit.

In consideration of these results, the EBB2 was confirmed as superior to the other two scales in terms of all the aspects examined here.

Insert Table 6 about here.

Validity of the Scales

Although we found that all three trait criteria of the EBB2 seem to discriminate successfully between participant performances, we were not certain whether this discrimination was based on the abilities that we intended to measure. To examine this validity issue, we investigated the relationships among the individual trait criteria of the three scales using a set of Pearson product-moment correlations called a multitrait-multimethod (MTMM) matrix (Crocker & Algina, 1986).

Correlation coefficients in boldface in Table 7 indicate convergent validity coefficients, which are correlations between measures of the same trait using different or the same measurement methods (i.e., the monotrait-heteromethod or monotrait-monomethod). Ideally, these correlation coefficients should be higher than the divergent validity coefficients (not in boldface), which are correlations between measures of different traits using different or the same measurement methods (i.e., the heterotrait-heteromethod or heterotrait-monomethod).

Insert Table 7 about here.

The overall results of the convergent validity coefficients, indicated in boldface, show that they were significant, and many, although not all, were higher than the divergent validity coefficients. For example, EBB1 Communicative Efficiency correlated with MT Communicative Efficiency more highly (.46) than it did with the other two MT criteria (.12 and .31). EBB1 Grammar & Vocabulary was also correlated with the Grammar & Vocabulary of the other scales more highly (.58 and .62) than with other criteria (.24 to .42). EBB1 Pronunciation correlated with both the EBB2 and MT Pronunciation criteria more highly than with the other divergent validity coefficients (i.e., .64 over .50 and .56; .39 over .25 and .10). These higher correlation coefficients were considered as evidence of convergent validity. Some convergent validity coefficients were relatively low due to various scale properties, as seen in Tables 3 to 5.

However, the EBB2 convergent validity coefficients were not very clear, as the correlations among the EBB2 criteria were higher (.73, 69, and .75) than those among the convergent validity coefficients (.31 to .64). Thus, to examine the validity of the EBB2 in relation to the other scales, we performed a PCA with the Equamax rotation. We regarded eigenvalues over 0.7 as components, following the suggestion of Jolliffe (1986) that Kaiser's criterion of over 1.0 is too strict, and that eigenvalues over 0.7 should be retained.

As shown in Table 8, the rotated component matrix revealed four components that explained as much as 82% of the total variance. The same or similar traits tended to load highly on the same components, which indicates that the individual criteria of each scale generally assess the intended constructs.

Insert Table 8 about here.

However, as seen in the MTMM matrix, we found that the situations of the EBB2 and other two scales were somewhat different. All the trait criteria of the EBB2 loaded highly on the first component (.72, .78, and .77) and were weakly loaded on the following three components separately. EBB2 Communicative Efficiency (.31) and its related traits, such as EBB1 Communicative Efficiency (.89), EBB1 Content (.85), and MT Communicative Efficiency (.49), loaded highly on the second component, whereas the EBB2 Grammar & Vocabulary criteria (.35) loaded on the third component with the other Grammar & Vocabulary criteria (.81 and .91) and MT Communicative Efficiency (.62). EBB2 Pronunciation (.48) loaded on the fourth component with MT

Communicative Efficiency (.40) and Pronunciation (.92). In contrast, the EBB1 and MT individual criteria were separated more clearly on different components, apart from MT Communicative Efficiency.

These results indicate that the EBB2 assessed speaking abilities more holistically than did the EBB1 and MT.

Practicality of the Scales

The final issue is the practicality of the scales. Using the questionnaire results, we compared rater perceptions of the practicality of the EBB2 and MT (see Appendix E). The raters reported that the amount of rater training was sufficient at one or two hours for each scale. Concerning Question (2), all raters were able to give scores by listening to a speech only once, using either the EBB2 or MT. Given that three out of five raters who used the EBB1 in the preliminary study could not accurately score a speech during a single listening to it only once (Hirai & Koizumi, 2008), the practicality of the EBB2 improved on that of the EBB1.

However, regarding Questions (3) and (4), four out of seven raters reported having spent less time on scoring performances using the MT, and five found it easier to use. As indicated in Question (5), a main reason for the ease of the MT was the difference in each scale's format. Raters 2 and 6 felt that it took time to decide "Yes" or "No" for each descriptor in the EBB2. In addition, Raters 1 and 7 felt it was easier to form judgments on the MT by looking at the five-point scales all together. Only two raters felt that the EBB2 was better and easier to use. Rater 3 mentioned that it was efficient and easy to go down the branches, and Rater 5 admitted that she was more careful in making decisions at each branching node.

Thus, overall, while the MT is slightly easier to use, the EBB2 helps in more careful scoring. In addition, compared with the EBB1, the EBB2 is practical enough to use for giving ratings during an actual two-minute performance.

DISCUSSION

Constructing a rating scale is an on-going process. The present study, as part of this process, compared three rating scales to determine which one was suitable for assessing performances in the Story Retelling Speaking Test (SRST), by seeking to answer four research questions (RQs).

First, regarding RQ1 (How many stories are needed to achieve sufficient reliability when the EBB1, EBB2, and MT rating scales are used for scoring?), the results revealed that the EBB2 needed only two stories. On the other hand, the MT could not overcome

rating inconsistency regarding the Grammar & Vocabulary criterion and still required three stories, as did the EBB1. Thus, by administering an SRST that contains only two stories and using the EBB2 for scoring, we were able to obtain reliable speaking scores.

Next, in answer to RQ2 (To what extent do the individual trait criteria of the EBB1, EBB2, and MT differ in terms of (a) the discrimination of the rating scale, (b) rater separation, (c) rater reliability, (d) variability in the ratings, and (e) rating scale properties?), overall, the EBB2 functioned best, and discriminated between good and poor performances more clearly and reliably. In addition, problematic wordings in the EBB1 (e.g., the influence of utterance length on the occurrence of grammatical errors in the Grammar & Vocabulary criterion and of the vague wording in the Pronunciation criterion) were found to be successfully solved by the modification of the descriptors in the EBB2. These results suggest the importance of carefully defining descriptors, because a poorly written boundary descriptor can influence the final score, especially if it appears early in the decision-making process.

The MT, in contrast, turned out to be the least discriminating of the three scales, and more raters rated inconsistently when they used it. The questionnaire revealed that raters found the MT easy to use because its format allowed them to see the full range of the scale at once. However, the ease of viewing all the descriptors at the same time seemed to cause raters to waver in their judgment and give fluctuating ratings across the five levels. EBB scales, on the other hand, require a yes-or-no decision to move on to the next level, so raters are forced to concentrate on only one descriptor at each node; it is relatively difficult to look at the descriptors of the other direction of the node. Therefore, the analytic format appears easier to use, but it makes consistent rating more difficult when compared to the EBB format.

To collect evidence of construct validity for the scales addressed in RQ3 (To what extent are the individual trait criteria of the EBB1, EBB2, and MT scales related to each other?), we created an MTMM matrix and generally detected high convergent validity coefficients, such as for the Grammar & Vocabulary criterion between the EBB1 and EBB2 and between the EBB1 and MT. In addition, the results of the PCA revealed that the same or similar traits tended to load highly on the same components. Thus, the criteria of each scale generally assess the intended constructs.

However, the correlation coefficients among the individual trait criteria of the EBB2 were higher than the convergent validity coefficients, and in the PCA, all three EBB2 criteria loaded highly on the first component. The first component can be interpreted as a general speaking ability that participants need to describe and express their opinions on what they have read.

One reason for this unidimensional tendency may be that each EBB2 criterion involving multiple aspects of speaking ability reduced its specificity as a trait criterion. For example, Communicative Efficiency in the EBB2 involves fluency, content, and coherency, which led to more holistic rating. In this regard, this tendency of the EBB2 may indicate its limited use as a diagnostic tool. However, for classroom use, the EBB2 may still provide students with useful feedback for the following two reasons. First, the EBB2 additionally showed the various dimensions that appeared in the results of the PCA. Second, because the score descriptions are related to classroom learning objectives, giving students the EBB2 descriptor sheet containing their score can be useful in helping them plan ways to raise their speaking ability to the next level.

The MT, on the other hand, did not show such a tendency, possibly because, either consciously or unconsciously, raters might rate each of the three criteria differently according to their impressions of the speech performance instead of visually considering each level of the MT criteria. Another reason might be the poor wording of the descriptors. Although we tried to make the MT comparable to the EBB2, there may be room for improvement regarding the wording of its descriptor.

Lastly, concerning RQ4 (What are the raters' perceptions of the practicality of the EBB1, EBB2, and MT scales?), the results of the questionnaire delivered to the raters revealed that the EBB2 was perceived as slightly less practical than the MT. However, when compared to the EBB1, the EBB2 was shown to be an improvement. Every rater was able to provide scores for individual criteria while listening to a speech only once, which had not been possible using the EBB1. Thus, it is clear that the improvement in practicality of the EBB2 was primarily due to the reduction in the number of criteria from four to three, through combining the Communicative Efficiency and the Content criteria.

Another point regarding practicality is the difference between Turner and Upshur's (1996, 2002) rubrics and our rubrics in the EBB2. In their rubrics, the first decision point, or node, occurs at the center of the criterion, separating the score into either the upper half or lower half of the score range. In other words, the rater must first decide if the score is in the range of 1 to 3 or 4 to 6. In our criteria, the first decision starts at one end of the scale; that is, the decision is made at a point of either 5 or lower or at a point of 1 or higher. Therefore, our rubrics require more decisions to reach the final score—a maximum of four decisions for a 5-point scale in contrast to a maximum of three decisions for a 6-point scale in their rubric. It is unclear which leads to more reliable ratings, as we created the EBB scale based on prominent features of the speech

performances. However, it will be more practical to put the first decision point at the center of the score range to reduce the overall number of decision points.

At the same time, we expect that rating behaviors will change to some degree since one rater in this study (Rater 5) was aware of the impact of selecting inaccurate directions at early branching nodes of an EBB scale and changed his rating style in response to the scale format. His behavior demonstrates that some raters are affected by the placement of decision points. In this regard, these raters' comments can provide us with useful information on which areas raters find difficult and for which they need further training.

Considering these results from various perspectives, the modification of the EBB1 was largely successful. That is, the EBB2 showed high reliability and practicality; a result which accords with Turner and Upshur's (1996) study. Therefore, the EBB2 seems best suited to the SRST.

Limitations and Pedagogical Implications

Although we were able to develop a better rating scale in this study, some limitations were found that may need further investigation. First, the sample size was relatively small. Having only 48 participants may have limited the variety of performance patterns of the test, which may in turn limit the generalization of the study results. For the same reason, the relatively small number of raters and the use of only a questionnaire method may limit the results regarding the relative ease of use of the different rating scale types. With an increase in the number of raters and the use of a verbal protocol, we may be able to more precisely identify the factors that cause rating behaviors to differ across rating scale types. Third, the MT did not function well. This might have been due to the poor wording of the MT descriptors or because the MT format may not be suitable when the descriptors of a scale criterion mention different traits, thereby giving raters the impression of disjointedness. Thus, the wording of the MT descriptor should be examined in order to make it more effective. Fourth, as mentioned at the end of the Discussion section, we need to find rationales regarding the place of the first decision point in an EBB criterion (i.e., either at the center of the criterion or not) in terms of the reliability, validity, and practicality of the scale.

At present, EBB scales are not commonly used, perhaps in part because the scales are considered as task-specific, so that it may take time to create EBB scales for each task. However, the present study demonstrates that, in the context of the SRST, the EBB scale can be used with different stories for more consistent scoring.

Furthermore, our study offers two cautionary notes regarding the creation and use of rating scales. First, scale format affects rating consistency. An MT scale seems to be highly practical, but owing to its ease of rating, unless descriptors are highly explicit, the judgments of raters tend to fluctuate. On the other hand, since an EBB scale constantly requires binary decisions to move on to the next level, raters tend to be more cautious about their ratings, which may consequently lead to more consistent scoring.

The second issue is the number of trait criteria in a scale and what these criteria measure. In order to provide a proper diagnostic assessment for participants, several trait criteria, each of which focuses on one trait, may be ideal. However, to make the scale easier to use in class, it may be preferable to reduce the number of criteria, perhaps to three in the case of five levels in each criterion, when assessing performances that are approximately two minutes long. In addition, if one criterion includes multiple traits, it may weaken the diagnostic profile of participants' speaking ability. We therefore need to consider the balance between the practicality, reliability, and validity of the scales that we choose for assessing speaking performance.

ACKNOWLEDGEMENT

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI, C) [grant number 19520477 and 23520744]. We would like to thank the LAQ reviewers for their thorough reviews and many valuable comments on our earlier version of this paper.

REFERENCES

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70,* 380-390.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12,* 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of rating scale and rater experience. *Language Assessment Quarterly 7,* 54-74.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Center for Advanced Research on Language Acquisition (2009). Evaluation. Retrieved from http://www.carla.umn.edu/assessment/VAC/Evaluation/rubrics/

types/traitRubrics.html

Center for Advanced Studies in Measurement and Assessment (University of Iowa, College of Education). (2009). Download GENOVA suite programs. Retrieved from http://www.education.uiowa.edu/casma/GenovaPrograms.htm

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* New York: Harcourt College Publishers.

Fulcher, G. (2003). *Testing second language speaking.* Essex, U.K.: Pearson Education Limited.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29,* 395-420. doi: 10.1177/0265532211428317

Hirai, A., & Koizumi, R. (2008). Validation of an EBB scale: A case of the Story Retelling Speaking Test. *JLTA (Japan Language Testing Association) Journal, 11,* 1-20.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly, 6,* 151-167.

Honda, T. (2007). *Assessment of speaking performance in Japanese junior high school EFL classes: Task types and good task combinations.* Unpublished master's thesis, Tokyo Gakugei University, Japan.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Jolliffe, I. T. (1986). *Principal Component Analysis.* New York: Springer-Verlag.

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5,* 1-36. English Language Institute, University of Michigan. Retrieved from http://www.lsa.umich.edu/UMICH/eli/Home/Research/Spaan%20Fellowship/pdfs/SpaanV520007KnochFinal.pdf

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26,* 275-304.

Koizumi, R., & Hirai, A. (2010). Exploring the quality of the Story Retelling Speaking Test: Roles of story length, comprehension questions, keywords, and opinions. *ARELE (Annual Review of English Language Education in Japan), 21,* 211-220.

Lado, R. (1964). *Language teaching: A scientific Approach.* New York: McGraw-Hill.

Linacre, J. M. (2009a). Facets: Rasch-measurement computer program (Version 3.65.0) [Computer software]. Chicago: MESA Press.

Linacre, J. M. (2009b). A user's guide to Facets: Program manual 3.66.0. Chicago: MESA Press. Retrieved October, 20, 2009 from http://www.winsteps.com/a/facets.pdf

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 833-855). Mahwah, NJ: Lawrence Erlbaum.

Luoma, S. (2004). *Assessing speaking.* Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. Essex, U.K.: Addison Wesley Longman Limited.

Ministry of Education, Culture, Sports, Science and Technology. (n.d.). *Section 13 English.* Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/__icsFiles/afieldfile/2011/04/11/1298353_24.pd

National Institute for Educational Policy Research of Japan. (2007). *The investigation on the special project 'English speaking.'* Retrieved from http://www.nier.go.jp/kaihatsu/tokutei_eigo/05002051033004000.pdf

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension, 37*(2). Retrieved from http://www.joe.org/joe/1999april/tt3.php.

Society for Testing English Proficiency. (2012). *Eiken: Test in Practical English Proficiency.* Retrieved from http://stepeiken.org/

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.). *The language testing cycle: From inception to washback* (pp. 55-79). Melbourne, Australia: Applied Linguistics Association of Australia.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36,* 49-70.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12.

Appendix A. Story Retelling Speaking Test

Read the story silently within two minutes.

Story 1

Kenji goes to school by train. One morning he was very sleepy. After he left the station, he remembered that he left his bag in the train. Some textbooks, a box lunch and a dictionary were in the bag. At school he telephoned the lost-and-found office of the station to ask about the bag. But “We don’t have your bag” was the answer. He was shocked. He returned home and told his mother about it. His mother said, “You are lucky. A kind man brought your bag to the house. He found your name and address on it.”

After the signal, read each question aloud and answer it in English.
Q1: Where did Kenji leave his bag?
Q2: What was there in the bag?
Q3: Why was Kenji lucky?

----------<Next page>----------

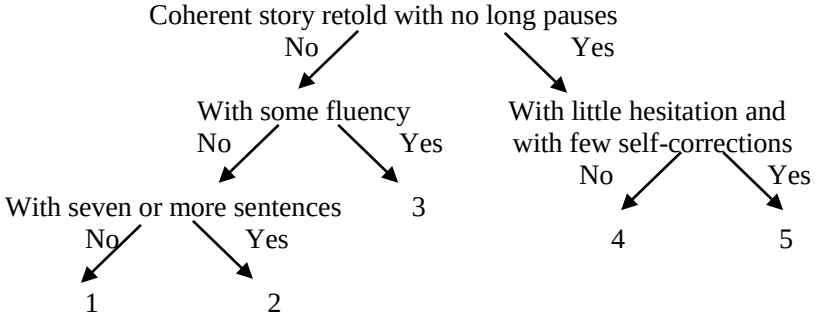
Retell as much of the story as you can in English in two and half minutes. You can look at the keywords while you are retelling. At the end of your retelling, be sure to include your opinions about the story.

Keywords:

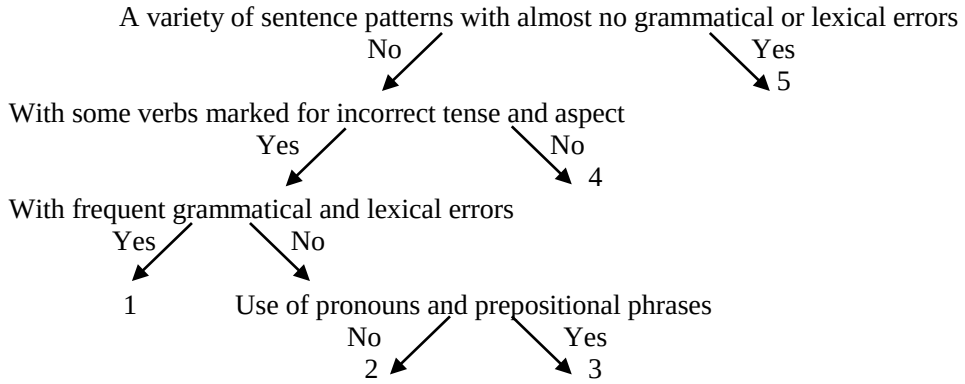
Kenji, train, bag, mother

Appendix B. EBB1 scale for SRST

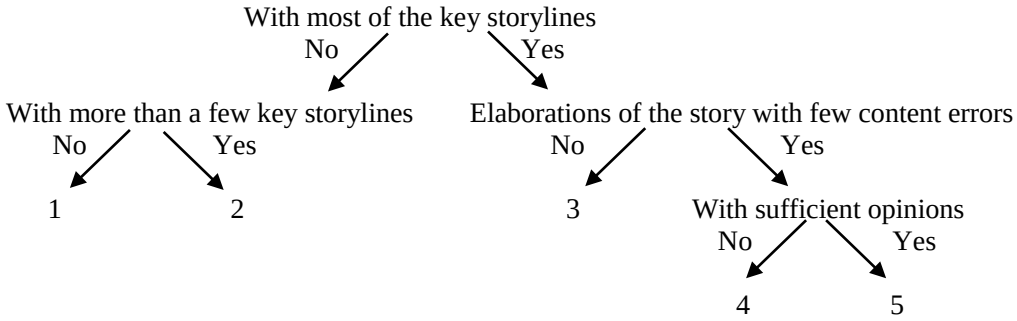
1. Communicative Efficiency



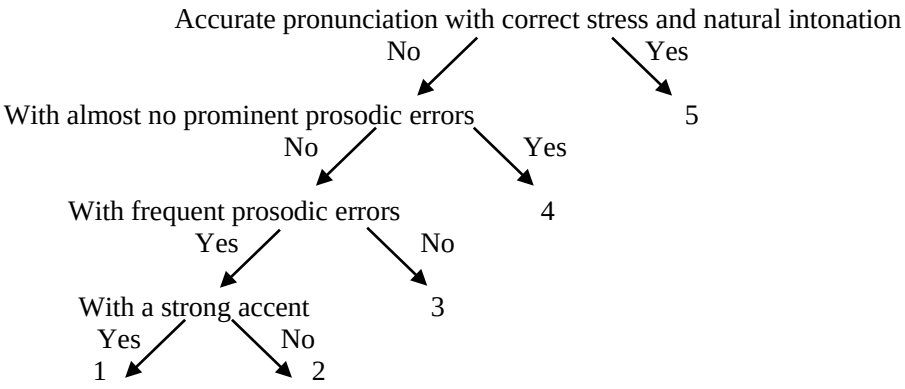
2. Grammar & Vocabulary



3. Content

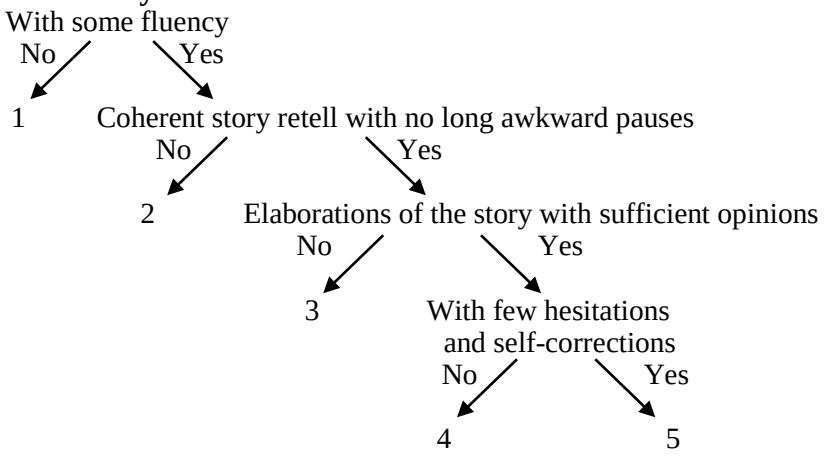


4. Pronunciation

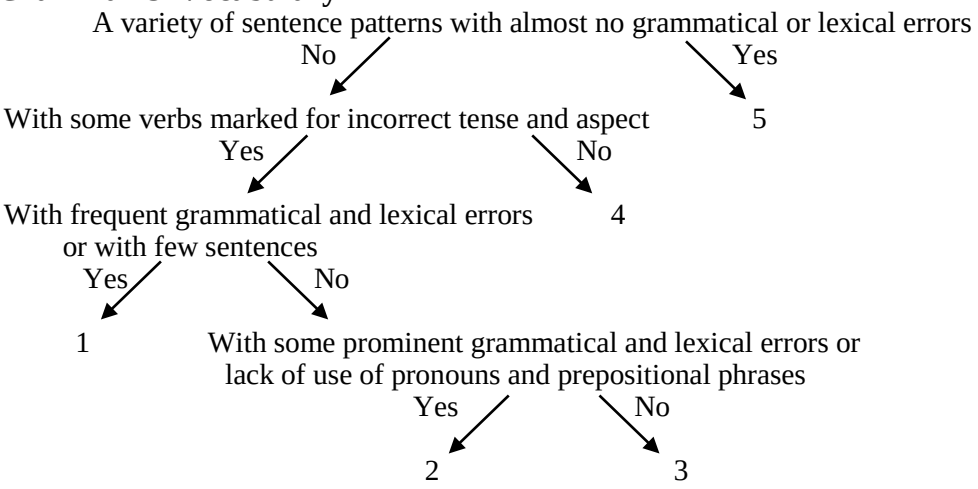


Appendix C. EBB2 scale for SRST

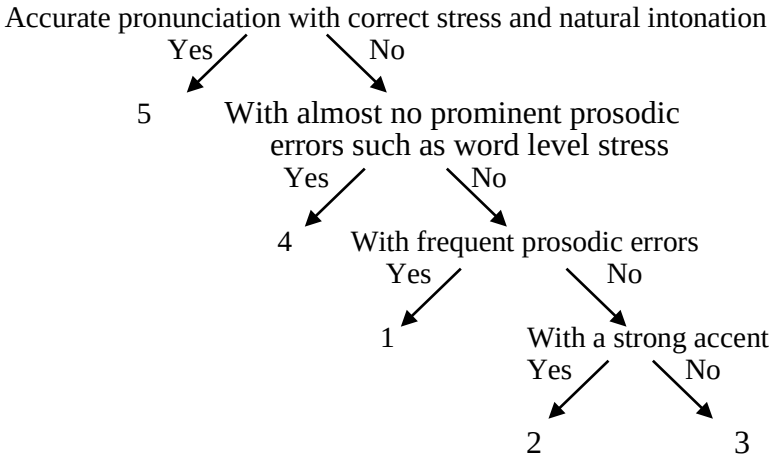
1. Communicative Efficiency



2. Grammar & Vocabulary



3. Pronunciation



Note. Examples of "a variety of sentence patterns" in the Grammar & Vocabulary criterion include patterns of a subject and a general verb (e.g., *Kenji goes to school*), a subject and a verb *be* (e.g.,

Some textbooks were in the bag), a main clause with a subordinate clause (e.g., *After he left the station, he remembered*), an infinitive (e.g., *He called to ask about the bag*), and a passive voice (e.g., *He was shocked*). To judge whether students can produce a variety of sentence patterns, texts for the SRST would need to include a variety of sentence patterns.

In the Content criterion, any opinions are considered adequate as long as they are related to the story retold.

Appendix D. Multiple-trait (MT) scale for SRST

5. Excellent; 4. Very Good; 3. Good; 2. Fair; 1. Poor

Aspect	Descriptor
Communicative Efficiency	5. Coherent and elaborate story retell with sufficient opinions, with few hesitations and self-corrections. 4. Coherent and elaborate story retell with sufficient opinions, but with some hesitations and self-corrections. 3. Few elaborations of the story or not sufficient opinions with no long awkward pauses. 2. Lack of coherency or with some long awkward pauses. 1. Little fluency.
Grammar & Vocabulary	5. A variety of sentence patterns with almost no grammatical or lexical errors. 4. A few grammatical and lexical errors but most verbs marked for correct tense and aspect. 3. Some verbs marked for incorrect tense and aspect, but correct use of pronouns and prepositional phrases. 2. Some prominent grammatical and lexical errors, or few use of pronouns or prepositional phrases. 1. Frequent grammatical and lexical errors or only a few sentences.
Pronunciation	5. Accurate pronunciation with correct stress and natural itonation. 4. Almost no prominent prosodic errors, but with some inaccurate pronunciation, incorrect stress, or unnatural intonation. 3. Some prosodic errors and with no strong accent. 2. Some prosodic errors and with a strong accent. 1. Frequent prosodic errors.

Appendix E.

Raters' Responses Toward EBB2 and MT Scales

Rater	(1) Was the rater training sufficient? (hours)	(2) Were you able to score a speech while listening to it once?		(3) Which required more time?	(4) Which was easier to use?	(5) State reasons for your answers to (3) and (4).[a]
		EBB2	MT			
R1	Sufficient (1.5h)	Yes	Yes	EBB2	MT	The level of the MT scale became more familiar as I went through it more, and thus it became easier to judge while listening.
R2	Sufficient (1h)	Yes	Yes	EBB2	MT	It is time consuming to follow down the Yes's and No's.
R3	Sufficient (EBB: 2h MT: 1h)	Yes	Yes	Same	EBB2	It was efficient and easy to just go down the branches.
R4	Sufficient (EBB: 2h MT: 1h)	Yes	Yes	Same	MT	Both took a lot of time at first, but as I got done more, the more I got used to them.
R5	Sufficient (EBB: 2h MT: 1h)	Yes	Yes	EBB2	EBB2	I was more careful in making decisions at each branching nodes, because each decision on which branch to choose could make a big difference in the final score.
R6	Sufficient (EBB: 2h MT: 1h)	Yes	Yes	EBB2	MT	MT was easier once you memorized the scales. EBB sometimes gave you hard time making decisions at each node.
R7	Sufficient (EBB: 2h MT: 1h)	Yes	Yes	Same	MT	EBB requires more time to understand what is to be evaluated in the big picture. In the MT, on the other hand, it is easier for the rater to get a bird's eye view of what is being assessed by looking at the 5-point scale.

Note. [a]Answered in Japanese and translated later.

Table 1

Estimated Variance Component and Proportion of Variance Explained in EBB1, EBB2, and MT Scales

EBB1	Persons (*p*)	Stories (*s*)	Residual (*ps, e*)
Communicative Efficiency	0.772 (74.09%)	0.011 (1.05%)	0.259 (24.86%)
Content	0.700 (64.17%)	0.071 (6.48%)	0.320 (29.36%)
Grammar & Vocabulary	0.445 (45.95%)	0.000* (0.00%)	0.524 (54.05%)
Pronunciation	0.837 (76.90%)	0.001 (0.08%)	0.250 (23.01%)
EBB2	Persons (*p*)	Stories (*s*)	Residual (*ps, e*)
Communicative Efficiency	0.559 (62.07%)	0.010 (1.15%)	0.331 (36.78%)
Grammar & Vocabulary	0.859 (72.45%)	0.003 (0.29%)	0.323 (27.26%)
Pronunciation	1.015 (82.09%)	0.001 (0.08%)	0.220 (17.83%)
MT	Persons (*p*)	Stories (*s*)	Residual (*ps, e*)
Communicative Efficiency	0.558 (60.28%)	0.007 (0.79%)	0.360 (38.94%)
Grammar & Vocabulary	0.421 (50.21%)	0.009 (1.08%)	0.408 (48.71%)
Pronunciation	0.402 (60.84%)	0.000* (0.00%)	0.258 (39.16%)

Note. *Negative variance was set to zero.

Table 2

Phi Coefficient (Φ) in Decision Studies of the EBB1, EBB2, and MT Scales for Each Criterion (p x s Design)

EBB1	1 story	2 stories	3 stories	4 stories
Communicative Efficiency	**.741**	.851	.896	.920
Content	.642	**.782**	.843	.878
Grammar & Vocabulary	.460	.630	**.718**	.773
Pronunciation	**.769**	.869	.909	.930
EBB2	1 story	2 stories	3 stories	4 stories
Communicative Efficiency	.621	**.766**	.831	.867
Grammar & Vocabulary	**.725**	.840	.888	.913
Pronunciation	**.821**	.902	.932	.948
MT	1 story	2 stories	3 stories	4 stories
Communicative Efficiency	.603	**.752**	.820	.859
Grammar & Vocabulary	.502	.669	**.752**	.801
Pronunciation	.608	**.757**	.823	.861

Note. Boldface = phi coefficient over .70 in the smallest number of stories.

Table 3

Rating Scale Statistics for Communicative Efficiency

	EBB1			
	CE	Content	EBB2	MT
(a) Person discrimination				
Person separation ratio:	2.41	2.20	2.47	2.33
(b) Rater separation				
Rater separation ratio:	3.07	3.38	1.22	2.53
(c) Rater reliability				
Rater point biserial:	.40	.39	.31	.29
Exact agreement ratio:	60.0%	56.2%	66.2%	50.0%
(d) Variation in ratings				
Rater misfit (%):	1/6[a] (16.7)	1/6 (16.7)	0/9 (0.0)	1/9 (11.1)
Rater overfit (%):	1/6 (16.7)	0/6 (0.0)	1/9 (11.1)	2/9 (22.2)
(e) Scale properties	Count (%) [logit]	Count (%) [logit]	Count (%) [logit]	Count(%)[logit]
Level 1	21 (8%) [-1.44]	8 (3%) [-1.00]	11 (4%) [-2.15]	18(7%) [-1.77]
Level 2	38 (14%) [-1.15]	61(22%)[-0.23]	94 (35%) [-1.11]	78(31%) [-1.04]
Level 3	86 (32%) [-0.10]	70(26%) [0.41]	107(39%) [-0.30]	99(40%) [-0.28]
Level 4	114 (42%) [0.29]	95(35%) [0.77]	47 (17%) [1.09]	40 (16%) [0.58]
Level 5	13 (5%) [2.54]	38(14%) [2.31]	13 (5%) [3.74]	13 (5%) [2.75]

Note. [a]1/6 reads one out of six raters.

Table 4

Rating Scale Statistics for Grammar & Vocabulary

	EBB1	EBB2	MT
(a) Person discrimination			
Person separation ratio:	1.85	2.71	2.01
(b) Rater separation			
Rater separation ratio:	0.00	0.82	1.91
(c) Rater reliability			
Rater point biserial:	.30	.47	.23
Exact agreement ratio:	57.5%	51.2%	41.1%
(d) Variation in ratings			
Rater misfit (%):	0/6 (0.0)	0/9 (0.0)	2/9 (22.2)
Rater overfit (%):	1/6 (16.7)	1/9 (11.1)	0/9 (0.0)
(e) Scale properties	Count (%) [logit]	Count (%) [logit]	Count (%) [logit]
Level 1	31 (11%) [-0.73]	41 (15%) [-2.22]	33 (13%) [-1.96]
Level 2	8 (3%) [-0.35]	72 (26%) [-1.33]	55 (22%) [-1.47]
Level 3	98 (36%) [-0.24]	92 (34%) [-0.42]	102 (41%) [-0.88]
Level 4	112 (41%) [0.32]	48 (18%) [0.88]	53 (21%) [0.29]
Level 5	23 (8%) [1.75]	19 (7%) [2.76]	5 (2%) [2.66]

Table 5

Rating Scale Statistics for Pronunciation

	EBB1	EBB2	MT
(a) Person discrimination			
Person separation ratio:	2.92	2.91	2.44
(b) Rater separation			
Rater separation ratio:	3.63	1.55	2.34
(c) Rater reliability			
Rater point biserial:	.51	.43	.33
Exact agreement ratio:	50.0%	66.2%	39.3%
(d) Variation in ratings			
Rater misfit (%):	0/6 (0.0)	0/9 (0.0)	2/9 (22.2)
Rater overfit (%):	1/6 (16.7)	1/9 (11.1)	1/9 (11.1)
(e) Scale properties	Count (%) [logit]	Count (%) [logit]	Count (%) [logit]
Level 1	24 (9%) [-0.74]	23 (8%) [-2.39]	12 (5%) [-1.84]
Level 2	16 (6%) [-0.72]	87 (32%) [-1.05]	78 (31%) [-1.38]
Level 3	88 (32%) [-0.26]	94 (35%) [-0.20]	102 (41%) [-0.82]
Level 4	116 (43%) [0.46]	46 (17%) [0.81]	53 (21%) [0.48]
Level 5	28 (10%) [2.57]	22 (8%) [3.49]	3 (1%) [3.71]

Table 6

Rating Scale Statistics for All Criteria

	EBB1	EBB2	MT
(a) Person discrimination			
Person separation ratio:	3.60	3.94	2.96
(b) Rater separation			
Rater separation ratio:	3.72	2.32	3.05
(c) Rater reliability			
Rater point biserial:	.26	.29	.18
Exact agreement ratio:	55.9%	61.2%	43.5%
(d) Variation in ratings			
Rater misfit (%):	0/6 (0.0)	1/9 (11.1)	3/9 (33.3)
Rater overfit (%):	0/6 (0.0)	0/9 (0.0)	0/9 (0.0)
(e) Scale properties	Count (%) [logit]	Count (%) [logit]	Count (%) [logit]
Level 1	84 (8%) [-0.86]	75 (9%) [-2.23]	63 (8%) [-1.82]
Level 2	123 (11%) [-0.69]	253 (31%) [-1.20]	211 (28%) [-1.27]
Level 3	342 (31%) [-0.05]	293 (36%) [-0.35]	303 (41%) [-0.63]
Level 4	437 (40%) [0.46]	141 (17%) [0.89]	146 (20%) [0.49]
Level 5	102 (9%) [2.15]	54 (7%) [3.24]	21 (3%) [2.69]

Table 7

Correlation Coefficients for the Criteria of EBB1, EBB2, and MT scales

	EBB1				EBB2			MT		
	CE	Con	GV	Pro	CE	GV	Pro	CE	GV	Pro
EBB1_CE	--	**.818****	.339*	.474**	**.442****	.558**	.481**	**.458****	.119	.309*
EBB1_Con		--	.282	.524**	**.485****	.600**	.580**	**.482****	.145	.413**
EBB1_GV			--	.364*	.241	**.577****	.260	.417**	**.623****	.294
EBB1_Pro				--	.503**	.560**	**.637****	.250	.104	**.392****
EBB2_CE					--	.730**	.686**	**.434****	.192	.280
EBB2_GV						--	.747**	.392**	**.313***	.425**
EBB2_Pro							--	.271	.059	**.568****
MT_CE								--	.639**	.433**
MT_GV									--	.333*
Mean	3.219	3.276	3.365	3.344	2.888	2.721	2.802	2.831	2.823	2.789
SD	0.889	0.848	0.745	0.923	0.801	0.970	1.034	0.805	0.723	0.683

Note. *p < .05. **p < .01. N = 48. CE = Communicative Efficiency; GV = Grammar & Vocabulary; Con = Content; Pro = Pronunciation.

Table 8

Rotated Component Matrix of EBB1, EBB2 and MT

	Component				Communalities[a]
Scale	1	2	3	4	Extraction
EBB1_CE	.277	**.892**	.116	.088	.894
EBB1_Con	**.317**	**.847**	.072	.257	.888
EBB1_GV	**.374**	.095	**.807**	-.025	.801
EBB1_Pro	**.694**	.282	.062	.248	.626
EBB2_CE	**.720**	**.309**	.157	.160	.664
EBB2_GV	**.782**	**.323**	**.348**	.166	.864
EBB2_Pro	**.770**	.257	-.044	**.479**	.890
MT_CE	-.042	**.493**	**.624**	**.397**	.792
MT_GV	-.044	.014	**.913**	.235	.891
MT_Pro	.193	.111	.179	**.917**	.924
Variance accounted (%)	25.61	21.23	20.77	14.73	Total 82.34%

Note. Loadings over .03 are boldfaced. [a]Rotation Sums of Squared Loadings for each variable.